과제 2 기상 특성에 따른 안개 발생 진단

2-Stage 모델을 활용한 안개 발생 예측

접수 번호	240274	
팀명	갈라진 하늘	
팀원	조성우	김다민
	이정환	정승민

목차

1. 과제 개요

1.1 시정거리와 안개 예측

안개는 상대습도, 풍속, 기온 등의 다양한 기상 변수로 인해 수평 시정거리가 1km 미만으로 감소하는 현상이다. 본 과제는 기상 데이터 기반으로 안개 발생을 예측하는 것을 목표로 한다. 안개는 교통사고나 항공기 운항 지연 및 취소 등의 피해를 초래할 수 있다. 따라서, 안개 발생을 정확히 예측하면 이러한 사회 문제를 개선하는 데 기여할 수 있다.

1.2 분석의 어려움과 방향성

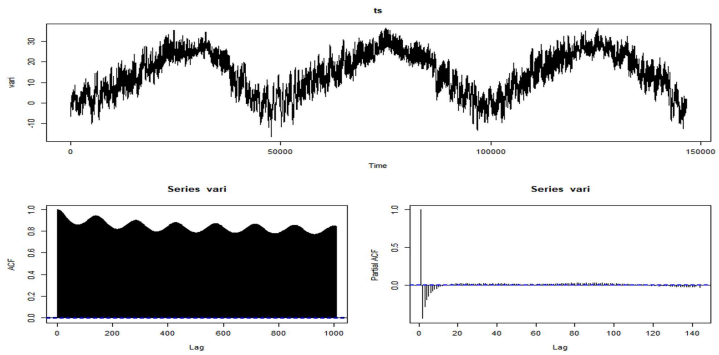
안개 발생 예측은 시정현천계로 측정한 변수만으로는 어렵다. 선행연구에서는 1시간 뒤 예측을 목표로 하여 시간 해상도가 낮았지만[1], 본 과제는 10분 단위 예측을 목표로 하여 시간 해상도가 매우 높아 안개 예측이 더욱 어려우며 증발 과정, 냉각 과정 등 복잡한 원인과 복사안개, 해양안개 등 종류도 다양하여 예측하기가 어렵다.[2]

또한, 안개가 발생하지 않은 시점이 발생 시점보다 절대적으로 많아 종속 변수인 클래스의 불균형 문제가 발생한다. 기존 연구는 이를 해결하기 위해 오버 샘플링과 언더 샘플링 기법을 사용했지만[1], 본 연구는 도메인 지식을 활용한 파생 변수와 Hurdle 모델에서 영감을 받은 2-Stage 모델링을 통해 이 문제를 해결하고자 한다.

2. 데이터 EDA

2.1 시계열 특성

제시된 데이터는 다변량 시계열 데이터이다. 모든 수치형 변수를 TS(Time series) plot과 ACF로 시각화한 결과, month와 time에서 비롯된 계절성이 존재하는 비정상 시계열성을 확인했다. 또한 10분 단위 데이터로 매우 먼 시점까지 lag이 존재했다.



[그림 1] ts 변수의 TS plot과 ACF,PACF

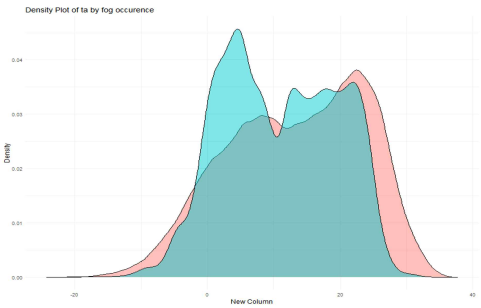
2.2 개별 변수 EDA

ws10_ms는 풍속을 의미하는 변수이다. 풍속은 시간대와 계절에 따라 변동하며, 풍속이 낮을 때 안개가 많이 발생하는 것을 확인할 수 있었다. 특히, 풍속이 10m/s를 넘어가는 경우에는 class 4에 해당하여 안개가 나타나지 않는 경향이 있었다.

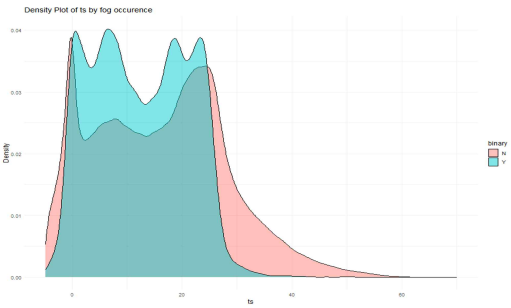
ws10_deg는 풍향을 의미한다. 육십분법으로 분석하기 어려워, 풍속과 결합하기 위해 직교좌표계로 변환하였다. 변환 후, 동서 방향으로 부는 바람의 세기를 나타내는 ws10_msx와 남북 방향으로 부는 바람의 세기를 나

1) 시정계 자료와 기계학습 기법을 이용한 지역 안개예측 모형 개발, 2021
2) 주요 안개 사례 분석을 통한 안개 분석기술과 예측방법, 2015

타내는 ws10_msy 변수를 얻을 수 있었다. ws10_msx와 ws10_msy 변수의 분포를 확인한 결과, 특별한 경향성은 확인되지 않았으며 두 변수 간의 상관관계도 0에 가까웠다. 한편, class에 대해서는 각각의 풍속이 클 때 class 4에 해당하는 관측치가 여전히 많았다.



[그림 2]



[그림 3]

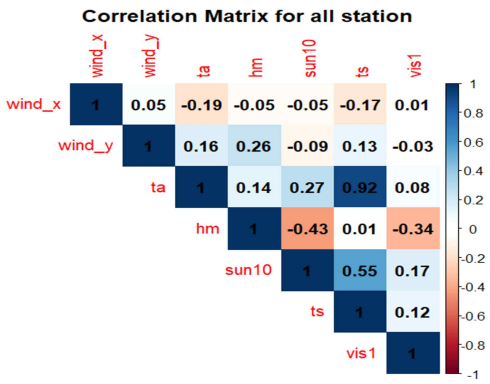
ta는 평균 기온을 의미하는 변수이다. 안개가 발생함(class 1~3)과 안개가 발생하지 않음(class 4)으로 나누면, ta가 0°C~10°C 구간에서 안개가 많이 발생한 것을 확인할 수 있었다(그림2). 다음으로 ts는 지면 온도를 의미하는 변수인데, 기본적으로 ta와 높은 양의 상관관계를 보인다. ts가 0°C~25°C 구간에서 대부분의 안개가 발생하는 것을 확인했다(그림3).

sun10은 10분 누적 일사량을 의미하는 변수이다. sun10에서 두 가지 이상치가 탐지되었다. 첫 번째는 sun10의 값이 1 이상인 경우이며, 두 번째는 해가 없는 시간대(20:00~05:00)에 sun10의 값이 0이 아닌 경우였다. 전자의 경우 해당 구간의 sun10 값을 결측치로 간주하였고, 후자의 경우 sun10의 값을 0으로 대체하였다.

hm은 상대습도를 의미하는 변수이다. 상대습도는 안개 생성에 가장 중요한 변수로, 상대습도가 낮을 때는 안개가 발생하지 않는 특징을 보인다. 또한, 상대습도는 기온에 영향을 받아 낮에는 하락하고 밤에는 상승하는 경향이 나타난다.

re는 강수의 유무를 의미하며, 분석 결과 re가 1일 때 습도가 전반적으로 높았다. 또한 class가 4보다 작을 때 re가 1인 관측치는 68.8%가 class 3에 해당했다. 즉, 강수가 있을 때 시정거리가 오히려 짧지 않았다.

2.3 변수 간 관계 EDA



[그림 4]

[그림4]는 전체 변수의 상관관계를 히트맵으로 표현한 것이다. 분석 결과, 습도(hm)와 일사량(sun10)은 음의 상관관계(-0.43)를, 일사량과 표면온도는 양의 상관관계(0.55)를 보였다. 이러한 정보는 직관과 일치하였다.

또한, 평균 온도(ta)와 표면 온도(ts) 사이에 높은 양의 상관관계(0.92)가 있음을 확인했다. 이는 일반적인 직

관과 일치하며, 상관관계 수치가 매우 높아 다중공선성 문제가 발생할 가능성도 있음을 시사한다.

상관계수가 1에 가까워 다중 공선성 문제가 우려되는 ta와 ts의 관계를 심층적으로 분석해 보았다. 주어진 데이터는 시계열성을 가지며, 관측소별 특성이 내재된 패널 데이터이다. 따라서 패널 선형회귀(PLM) 모델을 사용하여 R^2 값을 확인하고, 다중공선성의 지표인 VIF(분산 팽창 인자)를 점검하였다. 패널 선형회귀 모델을 검정한 결과, F-통계량은 $2.22e^{-16}$으로 0.05보다 작아 모델의 정당성을 확인했으며, R^2 값이 0.9382로 매우 높게 나타났다.

따라서 두 변수를 함께 사용하는 경우, VIF 값이 16.18로 10보다 커져 다중 공선성 문제가 발생할 수 있음을 확인하였으나, 데이터의 결측치가 많은 ts 변수의 보간에 ta를 사용하는 방안을 검토했다.

3. 결측치 보간 및 파생변수 생성

3.1 결측치 보간

주어진 데이터에서 train set에서는 ts(88,639), sun10(43,910), class(=vis1, 22,516), re(15,228), ws10_deg(5,910), ws10_ms(5,826), ta(3,867), hm(3,616) 순으로 결측치가 다량 발생했다. 한편 test set에서는 ts(7,643), re(4,503), ws10_deg(224), ws10_ms(224), sun10(115), ta(91), hm(57) 순으로 결측치가 발견되었다.

3.1.1 일사량(sun10) 결측치 보간

CA 지역에서 장기간(K년 8월 26일 00:10~K년 12월 23일 21:00) 연속된 결측치가 발견되었다. 이 구간의 일사량 값은 인접 지역인 CB 지역의 동일 기간 일사량 값으로 대체하였다.

3.1.2 변동계수에 따른 선형보간법

일사량의 결측치를 보간하기 위해 가장 먼저 선형보간법을 고려했다. 주어진 데이터는 시계열적 연속성을 가지고 있어, 시계열 선형보간법을 사용하였다. 이 방법은 결측 구간의 이전 시점과 이후 시점을 선형으로 연결하여 결측치를 채운다. Data leakage 문제를 방지하기 위해 train set에서만 선형보간을 적용했으며, 이항 변수인 re와 회귀 모델링이 필요한 ts는 선형보간에서 제외하였다.

전체 결측치를 일괄적으로 선형보간하면, 구간이 길어질 경우 정보 손실이 발생하고 모델 성능에 악영향을 미칠 수 있다. 따라서 선형보간을 적용할 결측 구간의 기준을 설정해야 한다. 선형보간을 적용할 결측 구간의 길이는 6으로 설정하였다. 이는 1시간 단위의 특수성과, 구간이 6보다 길어지는 결측 구간의 수가 이전보다 적다는 점을 고려하여 적절하다고 판단했다.

추가로, 변동계수를 사용하여 각 변수의 분산을 확인하고 선형보간 적용 구간을 조절하였다. 기존의 변동계수는 변수의 표준편차를 평균으로 나누어 구하지만, 시계열성을 고려해 표준편차 대신 1차 차분의 절대값 평균을 사용하였다. 변동계수는 ws10_deg(0.1768856), ws10_ms(0.1711653), sun10(0.1299905), hm(0.0157372), ta(0.0138127) 순으로 나타났다. 변동계수가 큰 상위 3개 변수(ws10_deg, ws10_ms, sun10)는 구간 길이를 6으로 고정하고, hm과 ta는 각 변수의 특성과 변동계수, 결측 구간 수를 고려해 각각 25와 19로 설정하여 보간하였다.

3.1.3 선형회귀를 이용한 ts 보간

앞서 언급한 것처럼, ts는 많은 결측치를 가지고 있으면서 ta와 높은 상관관계를 보인다. 또한, 패널 선형회귀 모형으로 적합했을 때 R^2값이 0.9382로 매우 높았다. 이를 근거로 나머지 변수를 사용한 선형 회귀모형을 사

용해 ts를 보간하였다. 시계열 데이터에 선형회귀를 사용해도 계수는 여전히 불편추정성(Unbiasedness)이 유지되기에 이는 적절한 결측치 대체 방법이다.

먼저, train 데이터의 ts를 보간하는 모델에는 stn_id를 사용하고, month와 time(hour) 변수를 삼각변환한 month_cos, month_sin, hour_cos, hour_sin 변수를 추가했다. ws10_msx와 ws10_msy 변수도 사용되었다. 그 결과, 모델의 R^2값은 0.9463으로 매우 높은 학습 성능을 보였다. 이를 학습한 뒤 ts의 결측치를 예측하여 보간하였다.

test 데이터의 ts를 보간하는 모델에는 stn_id를 적용할 수 없기 때문에, train에서 사용한 변수 중 stn_id 대신 FirstLetter(stn_id의 앞글자) 변수를 사용했다. 보간된 train 데이터에서 모델을 재적합한 결과, 모델의 R^2값은 0.9504로 매우 높았다. 이를 통해 test 데이터의 ts 결측치를 예측하여 보간하였다.

3.2 파생변수 생성

모델링을 위해 제작한 파생변수는 cosMD, sinMD, cosTM, sinTM, hmws10, half, dew, sup_dew, ta-ts, ta_lnhm으로, 각각 조사한 도메인적 지식이나 EDA에서 얻은 인사이트를 바탕으로 제작되었다.

cosMD, sinMD 변수는 Month Day, 즉 1월 1일부터 12월 31일까지의 날짜를 삼각변환하여 만들었으며, cosTM, sinTM 변수는Time Minute, 즉 0시 00분부터 10분 단위로 23시 50분까지의 시간을 삼각변환하여 만들었다. EDA에서 시계열 특성을 확인했기에, 주기성을 고려한 이 변수들로 계절성을 통제하도록 했다.

hmws10의 계산식은 풍속/(습도+$1e^{-7}$)이다. 습도가 낮고 풍속이 높은 경우에는 안개가 잘 발생하지 않으며, 반대로 습하고 바람이 불지 않는 경우 안개가 발생하기 쉽다는 관계를 모델에 반영할 수 있다.

half 변수는 4~10월, 11~3월을 구분한 이진 변수이다. 12개의 월을 4개의 계절이 아닌, 봄부터 초가을, 가을부터 늦봄으로 월을 분리했다. 이렇게 분리했을 때 ta와 같은 변수는 두 개의 정규분포로 나누어지는 모습을 보였다.

dew는 이슬점을 의미하는 변수이다. 이슬점은 공기의 수증기가 응결하여 물방울로 변하기 시작하는 온도인데, 공기의 온도가 이슬점에 근접하면 응결현상이 일어나게 되고 안개가 발생하기에 파생변수로 만들었다. 또한 기온이 이슬점보다 낮은 1, 아니면 0인 파생변수 sup_dew를 추가적으로 만들어, 기온이 이슬점보다 낮을 때 안개 발생 확률이 높아지는 특성을 모델에 반영하고자 했고, 이슬점 측정오차인 0.35를 반영했다.

ts-ta는 지면 온도와 대기의 평균 온도의 차이다. 지면 온도와 대기의 온도차로 인해 안개가 발생하는 것이 선행연구를 통해 밝혀졌기에 파생변수로 만들었다.[3)]

ta_lnhm은 기온과 로그 상대습도 간의 비율을 나타내는 변수이다. 안개는 높은 습도와 이슬점 아래에 기온에서 발생한다. 이를 산점도로 확인하여 보니 기온과 로그 상대습도가 일정한 비율일 때, 안개가 발생하는 경우가 있어 파생변수로 활용하였다. 로그 상대습도는 습도의 분포가 왼쪽으로 긴 꼬리를 가지고 있어 이를 보정해주기 위함이다.

4. 모델 선정 및 모델링

4.1 2-Stage 모델링

4.1.1 Hurdle 모델

안개 등급은 시정거리를 통해 분류하므로, 시정거리의 데이터 생성 과정을 고려하면 안개를 효과적으로 예측할 수 있다. 이러한 시정거리는 안개나 미세먼지 등 시정거리를 떨어트리는 요소들이 없는 경우에 20,000m 혹은 50,000m에 주로 분포했다. 즉, 안개가 발생하지 않을 때를 0으로 생각하면 시정거리는 영과잉 자료로 해석이

3) 시정계 자료와 기계학습 기법을 이용한 지역 안개예측 모형 개발, 2021`

가능하다.

영과잉 자료를 모델링하는 방식 중 하나로 hurdle 모델을 고려할 수 있다. hurdle 모델은 데이터 생성 과정에서 0이 발생하는 모델과 실제 값이 발생하는 모델을 분리하여 모델링하는 방식이다. 2014년 강수량 예측 연구에서 Poisson-Gamma 분포를 이용한 hurdle 모델을 활용한 바 있다.[4] 본 연구에서는 이러한 hurdle 모델을 변형하여 안개 유무와 개별 등급을 따로 예측하는 2-Stage 모델에 대해 다루려고 한다.

4.1.2 2-Stage 모델

Hurdle 모델에서 착안하여, 2-Stage 모델 아이디어를 도출하였다. Stage 1 모델의 목적은 시정거리 class의 대부분을 차지하는 class 4를 예측하는 것이 목적이고, Stage 2 모델의 목적은 남은 class 1, 2, 3을 예측하는 것이다. 이를 위해 Stage 1 모델은 모든 데이터에 대해 class 1, 2, 3과 class 4로 이진 분류를 수행한다. Stage 2 모델은 Stage 1 모델이 class 1, 2, 3으로 판단한 데이터에 대해 1부터 3까지의 class를 분류한다.

모델 학습은 다음과 같이 진행된다. Stage 1 모델 학습을 위해 모든 데이터의 시정거리를 class가 4이면 0, class가 1, 2, 3이면 1로 변환했다. 그리고 Stage 2 모델에서는 전체 데이터 중 class가 1, 2, 3인 데이터만 사용하여 학습했다. 특히, Stage 1에서 학습 후 F1 스코어를 최대화하는 threshold를 설정하였다.

학습이 완료되면, 전체 데이터를 Stage 1 모델로 예측하여 0과 1로 이진 분류한다. 모델이 1로 예측한 시점들만 Stage 2 모델을 통해 class 1, 2, 3로 분류한다. Stage 1 모델이 class 4를, Stage 2 모델이 class 1, 2, 3을 예측하므로 전 시점 및 class에 대한 예측이 가능하다.

지역별(A~E)로 같은 알고리즘 모델을 적용했으나, 지역별로 5개의 모델을 학습했기 때문에 각각 다른 학습 파라미터를 가진 별개의 모델이다. Stage 별로 다른 모델의 알고리즘을 사용했으며, 산정한 모델은 XGBoos, LightGBM, RandomForest 분류기이다.

4.2 모델 평가

모델 평가는 훈련 데이터와 검증 데이터로 분리하여 진행했다. 모델 및 하이퍼 파라미터 결정, 변수 선택을 이후 전체 데이터로 모델을 재학습하고, 홈페이지의 검증 결과로 제출하였다. 시계열 특성을 고려한 모델을 사용하지는 않았지만, 패널 데이터이므로 test의 L년과 가장 가까운 K년을 검증 데이터로, 나머지 I,J년을 훈련 데이터로 산정하였다.

4.2.1 최종 모델 하이퍼 파라미터 최적화

최종적으로 선택한 모델은 원본 데이터의 Feature에 파생변수 sinMD, cosMD, sinTM, cosTM, ta-ts, hmws10, dew, half, ta_lnhm, sup_dew를 사용하여 학습 및 예측을 진행한 2-Stage 모델이다.

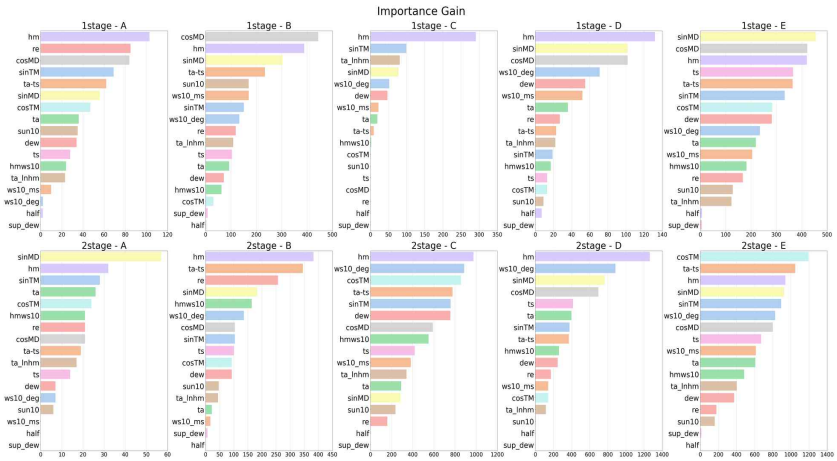
하이퍼 파라미터의 최적화는 Baysian Search 방식의 일종인 Tree-structured Parzen Estimator 알고리즘을 사용하여 진행했으며, 파이썬 패키지인 Optuna를 이용하여 구현했다. 하이퍼 파라미터 최적화의 목적은 검증 데이터의 CSI를 최대화하는 것으로 설정했다. 각 Stage에 어떤 모델을 사용할 것인지도 최적화 대상에 포함하여, 가장 좋은 조합을 찾게 했다. 하이퍼 파라미터 추천의 경우 Stage와 지역을 별개로 산정하였다. 다시 말해, 지역별 특성을 고려한 하이퍼 파라미터가 산정된다.

최적화를 진행한 결과 Stage 1, Stage 2 모두 LGBM 모델을 적용하는 것으로 홈페이지 검증 CSI 스코어 14.3%의 성능을 나타냈다. Stage 1의 threshold는 지역별로 A가 0.165, B가 0.155, C가 0.035, D가

4) A Poisson-Gamma Model for Zero Inflated Rainfall Data, Nelson Christopher Dzupire

0.08, E가 0.09였다.

4.2.2 최종 모델 평가



[그림5]

최종 모델의 변수 중요도를 살펴보면, 전반적으로 시간과 관련된 변수인 sinMD, cosMD, sinTM, cosTM가 중요도가 높은 모습을 보였다. 또한 hm, re, ta-ts, dew, ws10_deg 변수들도 높은 중요도를 보였다.

Metric	2-Stage Model
F1 Score	40.7522%
CSI Score	9.7786%
검증 CSI	14.3%

영과잉 데이터의 불균형 데이터이므로 정확도는 평가 기준에서 제외되었으며, 최종 모델의 Valid CSI 스코어는 9.7796%이었으며 검증 CSI는 14.3이었다.

5. 결론

5.1 분석 요약

안개 발생 예측을 목표로, 다양한 기상 변수를 포함한 패널 데이터를 사용했다. 데이터의 시계열 특성과 변수들 간의 관계를 탐색하고, 결측치를 보간하기 위해 선형보간법과 선형 회귀모형을 적용하여 데이터를 보간했다. 이후, 안개 발생 예측의 어려움과 클래스 불균형 문제를 해결하기 위해 파생변수를 생성하고, Hurdle 모델의 개념을 차용한 2-Stage 모델을 통해 안개의 유무와 1~3의 class를 단계적으로 예측했다.

5.2 분석의 의의 및 발전가능성

이 분석은 클래스 불균형 문제를 해결하기 위해 보편적인 샘플링 기반 방법들을 사용한 것이 아니라, Hurdle 모델의 개념을 차용하여 독자적인 2-Stage 모델을 개발하여 성능을 향상시켰다는 의의가 있다. 또한 결측치를 특정 값으로 일괄 대치하지 않고, 변동계수의 개념을 활용하여 변수마다 기준을 설정하기 위한 노력을 수행했다.

추후 분석에서 시계열적인 특성을 반영할 수 있으면, 누적 일사량이나 일교차와 같은 보다 강력한 파생변수를 추가적으로 생성할 수 있을 것으로 기대한다. 또한, test set에서의 결측치가 패턴이 없을 때에는 전시점으로 결측치를 대체하여 안정적인 분석을 수행할 수 있을 것이다. 마지막으로, 지역별 안개 발생 원인과 주요 발생 시기가 지리적인 요건에 의해 다르기 때문에 지역에 맞는 추가적인 정보가 제공된다면 모델의 성능 향상을 기대할 수 있을 것이다.